

ENFT: Efficient Non-Consecutive Feature Tracking for Robust Structure-from-Motion

Guofeng Zhang, Haomin Liu, Zilong Dong, Jiaya Jia, Tien-Tsin Wong, and Hujun Bao

Abstract—Structure-from-motion (SfM) largely relies on the quality of feature tracking. In image sequences, if disjointed tracks caused by objects moving in and out of the view, occasional occlusion, or image noise, are not handled well, the corresponding SfM could be significantly affected. This problem becomes more serious for accurate SfM of large-scale scenes, which typically requires to capture multiple sequences to cover the whole scene. In this paper, we propose an efficient non-consecutive feature tracking (ENFT) framework to match the interrupted tracks distributed in different subsequences or even in different videos. Our framework consists of steps of solving the feature ‘dropout’ problem when indistinctive structures, noise or even large image distortion exist, and of rapidly recognizing and joining common features located in different subsequences. In addition, we contribute an effective segment-based coarse-to-fine SfM estimation algorithm for efficiently and robustly handling large datasets. Experimental results on several challenging and large video datasets demonstrate the effectiveness of the proposed system.

Index Terms—Non-Consecutive Feature Tracking, Track Matching, Structure-from-Motion, Bundle Adjustment.

I. INTRODUCTION

Large-scale 3D reconstruction [34], [23], [14], [13], [8] finds many practical applications in, for example, Google Earth and Microsoft Virtual Earth. Recent work primarily relies on the SfM algorithms [16], [52], [2], [1], [48] to automatically estimate 3D features given the input of images or video collections.

Compared to images, videos contain denser geometrical and structural information, and are the main source of SfM in the movie and commercial industry. A common strategy for video SfM estimation is by employing feature point tracking [27], [38], [26], which takes care of the temporal relationship among frames. It is also a basic tool for solving a variety of computer vision problems, such as camera tracking, video matching, and object recognition.

In this paper, we discuss two critical problems for feature point tracking, which could seriously handicap SfM especially for large-scale scene modeling. We propose new methods to address them. One problem is the vulnerability of feature tracking to object occlusions, illumination change, noise, and large motion, which easily causes occasional feature drop-out and distraction. This problem makes robust feature tracking from long sequences challenging.

G. Zhang, H. Liu, Z. Dong and H. Bao are with the State Key Lab of CAD&CG, Zhejiang University. G. Zhang is also affiliated with Innovation Joint Research Center for Cyber-Physical-Society System, Zhejiang University. Email: {zhangguofeng, zldong, bao}@cad.zju.edu.cn, 172753015@qq.com.

J. Jia and T.-T. Wong are with The Chinese University of Hong Kong. Email: {leo.jia, tt.wong}@cse.cuhk.edu.hk

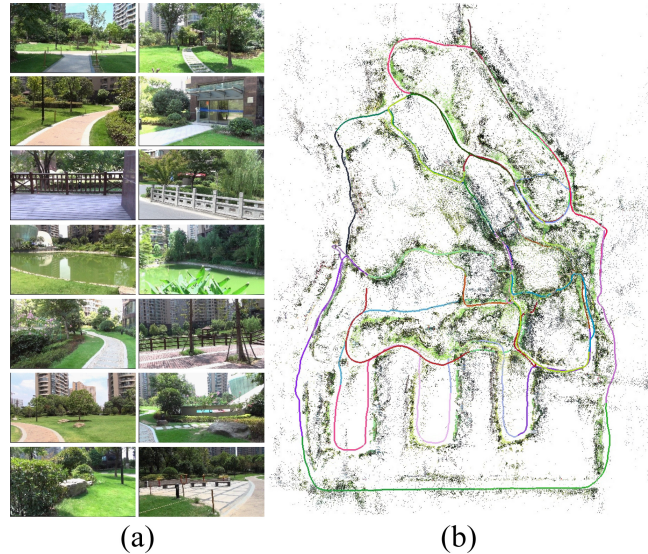


Fig. 1. A large-scale “Garden” example. (a) Snapshots of the input videos. (b) With the matched feature tracks, we register 3D points and camera trajectories in a large-scale unified 3D system. Camera trajectories are differently color-coded.

The other problem is the inability of sequential feature tracking to cope with feature matching over non-consecutive subsequences. A typical scenario is that the tracked object moves out and then re-enters the field-of-view, which yields two discontinuous subsequences containing the same object. Although there are common features in the two subsequences, they cannot be matched/included in a single track using conventional tracking methods. Addressing this issue can alleviate the drift problem of SfM, which in turn benefits high-quality 3D reconstruction as demonstrated in our experimental results. A naïve solution is to exhaustively search all features, which could consume much computation since many temporally far away frames simply share no content.

We propose an efficient non-consecutive feature tracking (ENFT) framework which can effectively address the above problems in two phases – that is, *consecutive point tracking* and *non-consecutive track matching*. We demonstrate their significance for SfM using a few challenging videos. *Consecutive point tracking* detects and matches invariant features in consecutive frames. A new matching strategy is proposed to greatly increase the matching rate and extend lifetime of the tracks. Then in *non-consecutive track matching*, by rapidly computing a match matrix, a set of disjoint subsequences with overlapping content can

be detected. Common feature tracks scattered over these subsequences can also be reliably matched.

Our ENFT method can help reduce estimation errors for long loopback sequences. Given limited memory, it is generally intractable to use global bundle adjustment to refine camera poses and 3D points for very long sequences. Iteratively applying local bundle adjustment is difficult to effectively distribute estimation errors to all frames. We address this issue by adopting a segment-based coarse-to-fine SfM estimation algorithm, which globally optimizes structure and motion only requiring limited memory.

Based on our ENFT algorithm and segment-based coarse-to-fine estimation scheme, we present a novel SfM system called ENFT-SfM, which can effectively handle long loopback sequences and even multiple sequences. Fig. 1 shows a challenging example containing 6 sequences with about 95,476 frames in total in a large-scale scene. Our system first splits them to 37 shorter sequences, then quickly computes many long and accurate feature tracks, efficiently estimates camera trajectories in different sequences and accurately registers them in a unified 3D system, as shown in Fig. 1(b). The whole process only takes about 90 minutes (excluding I/O) on a desktop PC, i.e., 17.7 FPS on average. Our supplementary video¹ contains the complete result.

A preliminary conference version paper was presented in [51]. In this manuscript, we have made a number of major modifications to improve robustness and efficiency. Particularly, we altered the second-pass matching by formulating it as minimizing an energy function incorporating two geometric constraints. We have developed an enhanced non-consecutive track matching algorithm, which can significantly reduce the matching time and robustly eliminate outliers. Finally, we proposed a novel segment-based coarse-to-fine SfM method, which can handle large sequence datasets with only limited memory.

II. RELATED WORK

We review feature tracking and large-scale SfM methods in this section.

A. Feature Matching and Tracking

For video tracking, sequential matchers are used for establishing correspondences between consecutive frames. Kanade-Lucas-Tomasi (KLT) tracker [27], [38], [50] is widely used for small baseline matching. Other methods detect image features and match them considering local image patches [32], [35] or advanced descriptors [26], [29], [28].

Both the KLT tracker and invariant feature algorithms depend on modeling feature appearance, and can be distracted by occlusion, similar structures, and noise. Generally, sequential matchers cannot match non-consecutive frames under image transformation. Scale-invariant feature

detection and matching algorithms [26], [3] are effective in matching images with large transformation. But they generally produce many short tracks in consecutive point tracking due primarily to the global indistinctiveness and feature dropout problems. In addition, invariant features are relatively sensitive to image distortion. Although variations, such as ASIFT [30], can improve matching performance under substantial viewpoint change, computation overhead significantly increases owing to exhaustive viewpoint simulation.

In this paper, we propose a novel two-pass matching method to address this problem. In [7], memory-based tracking method was used to extend feature trajectories, by matching each frame to its neighbors. However, if an object re-enters the field-of-view after a long period of time, the size of the neighboring windows has to be very large and the computation becomes expensive. Besides, this method cannot cope with multiple videos. Our method does not have this limitation, and the computation complexity is approximately linear to the number of processed frames.

There are methods using invariant features for object and location recognition in images/videos [41], [36], [18], [37], [19]. These methods typically use the bag-of-words technique to perform global localization and loop-closure detection in an image classification framework. For location recognition, Nistér and Stewénus [33] proposed using the feature descriptors to construct a vocabulary tree, and computing an appearance vector for each input image. Exhaustively comparing all image pairs is still time consuming for a long sequence. Cummins and Newman [9] proposed clustering similar images as a location, such that the computation can be reduced to only comparing the input image with previously visited locations. This method reduces the number of frames to be compared, but could perform less satisfyingly if consecutive frames have large overlaps.

In addition, these methods divide the location recognition and non-consecutive feature matching into two separated phases [24], [6], [10], [17]. Because the match matrix by bag-of-words only roughly reflects the match confidence, completely trusting it may lose many common features. In this paper, we introduce a novel strategy where the match matrix can be refined and updated along with non-consecutive feature matching. Our method can reliably and efficiently match the common features even with a coarse match matrix.

Engels et al. [11] proposed integrating wide-baseline local features with the tracked ones to improve SfM. The method creates small and independent submaps and links them via feature recognition. This approach also cannot produce many long and accurate point tracks. Short tracks are not enough for drift-free SfM estimation. In comparison, our method is effective in high-quality point track estimation. We also address the ubiquitous nondistinctive feature matching problem in dense frames. Similar to [15], we utilize track descriptors, instead of the feature descriptors, to reduce computation redundancy.

Wu et al. [49] proposed using dense 3D geometry infor-

¹<http://www.cad.zju.edu.cn/home/gfzhang/projects/tracking/featuretracking/featuretracking.wmv>

mation to extend SIFT features, which is called viewpoint invariant patch (VIP). In contrast, our method only uses sparse matches to estimate a set of homographies to represent scene motion, which also handles viewpoint change and is more general since geometry is not required.

B. Large-Scale Structure-from-Motion

The existence of large internet image datasets enlists the development of large-scale SfM. State-of-the-art methods can handle millions of images in a single PC in one day [13]. To this end, the large image sets are separated into a number of independent submaps, each of which is optimized independently. In [43], a partitioning approach was proposed to decompose large-scale optimization into multiple smaller and better-conditioned subproblems. In [5], local maps are built independently. They are stitched based on the hierarchical map approach.

Ni et al. [31] proposed an out-of-core BA for large-scale SfM. This method decomposes data into several submaps, all having their own local coordinate systems. They can be optimized in parallel. For global optimization, an out-of-core implementation is adopted. In [42], reconstruction efficiency is increased by selecting a skeletal image set and adding other images using pose estimation. Similarly, Konolige and Agrawal [22] selected a skeletal frame set in the frame-SLAM system, which only pre-edits and updates the motion of cameras, rather than positions of 3D features. Each skeletal view can actually be considered as a submap. The same scheme applies to iconic views [13], which are generated by clustering images with similar gist features. In our work, a segment-based scheme is adopted to estimate SfM, which first estimates SfM for each sequence independently, and then aligns the recovered submaps. Depending on estimation errors, we split each sequence to multiple segments, and perform segment-based refinement. This strategy avails handling large data, and quickly reduces estimation errors during optimization.

Another line of research is to improve large-scale BA, which is a core component of SfM. Agarwal et al. [1] pointed out that connectivity graphs of community photos are much less structured and accordingly presented an inexact Newton type BA algorithm, which offers decent performance for large-scale BA. To speed up large-scale BA, Wu et al. [48] utilized the computation power of multi-core CPUs or GPUs, and presented a parallel inexact Newton BA algorithm.

Note that most existing SfM approaches achieve reconstruction in an incremental way, which may risk drifting or local minima when dealing with large-scale image sets. Crandall et al. [8] proposed combining discrete and continuous optimization to yield a better initialization for BA. Discrete belief propagation is used to estimate camera parameters based on a Markov random field formulation. In the continuous step, Levenberg-Marquardt nonlinear optimization with additional constraints is used. This method is restricted to urban scenes, and assumes the vertical vanishing point can be detected for rotation estimation,

TABLE I
ENFT-SfM SYSTEM OVERVIEW.

1.	Consecutive point tracking (Section IV):
1.1	Match the extracted SIFT features between consecutive frames with descriptor comparison.
1.2	Perform the second-pass matching to extend track lifetime.
2.	Non-consecutive track matching (Section V):
2.1	Use hierarchical k-means to cluster the constructed invariant tracks.
2.2	Estimate the match matrix with the grouped tracks.
2.3	Detect overlapping subsequences and join the matched tracks.
3.	Segment-based coarse-to-fine SfM (Section VI):
3.1	Estimate the submap for each sequence.
3.2	Match the common tracks among different sequences, and then use them to estimate the similarity transformations for each submap.
3.3	Use segment-based SfM to refine the aligned submaps.
4.	[Optional] Feature Propagation with Camera Estimation:
4.1	Quickly propagate features from sampled frames to others.
4.2	Quickly estimate camera poses for remaining frames.

similar to that of [40]. It also needs to leverage the geotags contained in the collected images and takes complex discrete optimization. In contrast, our segment-based scheme can run on a common desktop PC with limited memory, even for large video datasets.

III. OUR APPROACH

Given a video sequence V with n frames, $V = \{I_t | t = 1, \dots, n\}$, our objective is to extract and match features in all frames in order to form a set of *feature tracks*. A feature track \mathcal{X} is defined as a series of feature points in images: $\mathcal{X} = \{\mathbf{x}_t | t \in f(\mathcal{X})\}$, where $f(\mathcal{X})$ denotes the frame set spanned by track \mathcal{X} . Each invariant feature \mathbf{x}_t in frame t is associated with an appearance descriptor $\mathbf{p}(\mathbf{x}_t)$ [26] and we denote all descriptors in a feature track as $\mathcal{P}_{\mathcal{X}} = \{\mathbf{p}(\mathbf{x}_t) | t \in f(\mathcal{X})\}$.

With the detected m features in all frames, finding matchable ones generally requires a large amount of comparisons even using the k-d trees; meanwhile it inevitably induces errors due to the fact that a large number of features make descriptor space hardly distinctive, resulting in ambiguous matches. So it is neither reliable nor practical to only compare the feature descriptors to form tracks. Our ENFT method has two main steps to address this issue. The algorithm is outlined in Table I.

For reducing computation, we extract one frame for every $3 \sim 5$ frames to constitute a new sequence and then perform feature tracking on it. In the consecutive tracking stage, we employ a two-pass matching strategy to extend the track lifetime. Then in the non-consecutive tracking stage, we match the common features in different subsequences. With the obtained feature tracks, we propose a novel segment-based SfM scheme to robustly recover the 3D structure and camera motion. Finally, if necessary, we propagate feature points from sampled frames to others. Since the 3D positions of these features have been computed, we can

Algorithm 1 Second-Pass Matching

- 1) Use the inlier matches to estimate a set of homographies $\{H_{t,t+1}^k | k = 1, \dots, N\}$ by Algorithm 2, and then use them to obtain a set of rectified images $\{\hat{I}_t^k | k = 1, \dots, N\}$.
 - 2) **for** each unmatched feature \mathbf{x}_t in I_t **do**

for $k = 1, \dots, N$ **do**

Find the best match \mathbf{x}_{t+1}^k by minimizing (1) with $H_{t,t+1}^k$.

end for

Find the best match \mathbf{x}_{t+1}^j among $\{\mathbf{x}_{t+1}^k | k = 1, \dots, N\}$ which minimizes $\sum_{\mathbf{y} \in W} |\hat{I}_t^k(\hat{\mathbf{x}}_t^k + \mathbf{y}) - I_{t+1}(\mathbf{x}_{t+1}^k + \mathbf{y})|$. Reject this match if it does not satisfy any of color and geometry constraints.

end for
-

quickly estimate the camera poses of remaining frames with the obtained 3D-2D correspondence.

IV. CONSECUTIVE TRACKING

For video sequences, feature tracks are typically obtained by matching features between consecutive frames. However, due to illumination change, repeated structure, noise, and large image distortion, features are easily dropped out or mismatched, resulting in breaking many tracks into shorter ones. In this section, we propose a *two-pass matching* strategy to alleviate this problem. The first-pass matching uses SIFT algorithm [26] to obtain high-confidence matches. In the second pass, tracks are extended with planar motion segmentation and constrained spatial search.

Without loss of generality, we discuss tracking a feature from I_t to I_{t+1} . It can be generalized to a track spanning multiple frames. An invariant feature in I_t is denoted as \mathbf{x}_t with descriptor $\mathbf{p}(\mathbf{x}_t)$. To determine if there is a corresponding feature \mathbf{x}_{t+1} with descriptor $\mathbf{p}(\mathbf{x}_{t+1})$ in I_{t+1} , we employ the 2NN heuristic proposed by [26]. However, this method is easily influenced by aforementioned problems. One example is shown in Fig. 2. Given an image pair, we detect 958 features in the first image, and 811 features in the second image. Only 53 features can be matched by descriptor comparison, as shown in Fig. 2(a). We thus propose a spatial search method to help retrieve more matches.

With a few high-confidence matches in neighboring frames (I_t, I_{t+1}) computed in the first step, we use the RANSAC algorithm [12] to estimate the fundamental matrix $F_{t,t+1}$ and remove outliers. For those unmatched features, it is possible to search for their correspondences along the conjugate epipolar line $l_{t,t+1}(\mathbf{x}_t) = F_{t,t+1}\mathbf{x}_t$. However, if significant image distortion or noise exists, key points may not be detected reliably, making SIFT features not perspective invariant. As shown in Fig. 2(b), directly searching correspondences along the epipolar lines by comparing SIFT descriptors does not increase much the

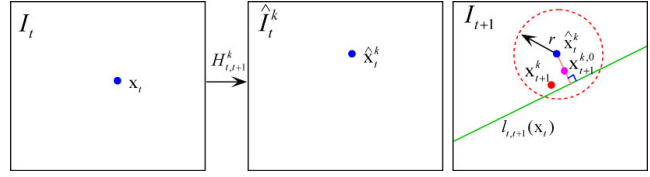


Fig. 3. Constrained spatial search with planar motion segmentation. Given homography $H_{t,t+1}^k$, we rectify I_t to \hat{I}_t^k such that $\hat{\mathbf{x}}_t^k \sim H_{t,t+1}^k \mathbf{x}_t$. Then we select the midpoint between $\hat{\mathbf{x}}_t^k$ and its projection to $l_{t,t+1}(\mathbf{x}_t)$ for initialization, and search the matched point by minimizing (1). The red dot \mathbf{x}_{t+1}^k is the result.

matches. Instead, we propose a segmentation-based method (sketched in Algorithm 1) to robustly identify missing matches.

We base our method on the observation that many feature points undergo similar motion. This allows for computing inlier matches to estimate a set of homographies $\{H_{t,t+1}^k | k = 1, \dots, N\}$, which represent possible local image transformation, as shown in Algorithm 2. We then rectify images with their homographies. This scheme is similar to that of [39] where a set of dominant scene planes are extracted to generate a piecewise planar depth map. For an unmatched feature in image I_t , if its transformation towards I_{t+1} is coincident with any of these homographies after rectification, a match in I_{t+1} can possibly be found. Incorrect homographies are unlikely to yield high-confidence matches. To handle illumination change, we estimate the global illumination variation $L_{t,t+1}$ between images I_t and I_{t+1} by computing the median intensity ratio between the matched features.

We first linearly scale image I_t with $L_{t,t+1}$, and then transform it with homography $H_{t,t+1}^k$ to obtain the rectified image \hat{I}_t^k . Correspondingly, \mathbf{x}_t in image I_t is rectified to $\hat{\mathbf{x}}_t^k$ where $\hat{\mathbf{x}}_t^k \sim H_{t,t+1}^k \mathbf{x}_t$ in \hat{I}_t^k . If $\hat{\mathbf{x}}_t^k$ largely deviates from the epipolar line (i.e., $d(\hat{\mathbf{x}}_t^k, l_{t,t+1}(\mathbf{x}_t)) > \tau_e$), we reject $H_{t,t+1}^k$ since it does not describe the motion of \mathbf{x}_t well. For each remaining $H_{t,t+1}^k$, we track \mathbf{x}_t to \mathbf{x}_{t+1}^k by minimizing the matching cost:

$$S_{t,t+1}^k(\mathbf{x}_{t+1}^k) = \sum_{\mathbf{y} \in W} \|\hat{I}_t^k(\hat{\mathbf{x}}_t^k + \mathbf{y}) - I_{t+1}(\mathbf{x}_{t+1}^k + \mathbf{y})\|^2 + \lambda_e d(\mathbf{x}_{t+1}^k, l_{t,t+1}(\mathbf{x}_t))^2 + \lambda_h \|\hat{\mathbf{x}}_t^k - \mathbf{x}_{t+1}^k\|^2, \quad (1)$$

where $\hat{\mathbf{x}}_t^k + \mathbf{y}$ are the points in the window W centered at $\hat{\mathbf{x}}_t^k$. The last two terms encourage \mathbf{x}_{t+1}^k to be along the epipolar line and obey the homography transformation respectively. The corresponding weights are $\lambda_e = |W| \sigma_c^2 / \sigma_e^2$ and $\lambda_h = |W| \sigma_c^2 / \sigma_h^2$, where σ_c , σ_e , and σ_h account for the uncertainty of intensity, epipolar geometry and homography transformation respectively. In our experiments, these values are by default $\sigma_c = 0.1$ (for intensity values normalized to $[0, 1]$), $\sigma_e = 2$ and $\sigma_h = 10$. Note that σ_h is relatively large because we do not require the points to strictly lie on the same plane. As long as the point is near the plane, H_k can cancel the major distortion and provide a better matching condition.

Similar to KLT tracking, we solve for $S_{t,t+1}(\mathbf{x}_{t+1}^k)$

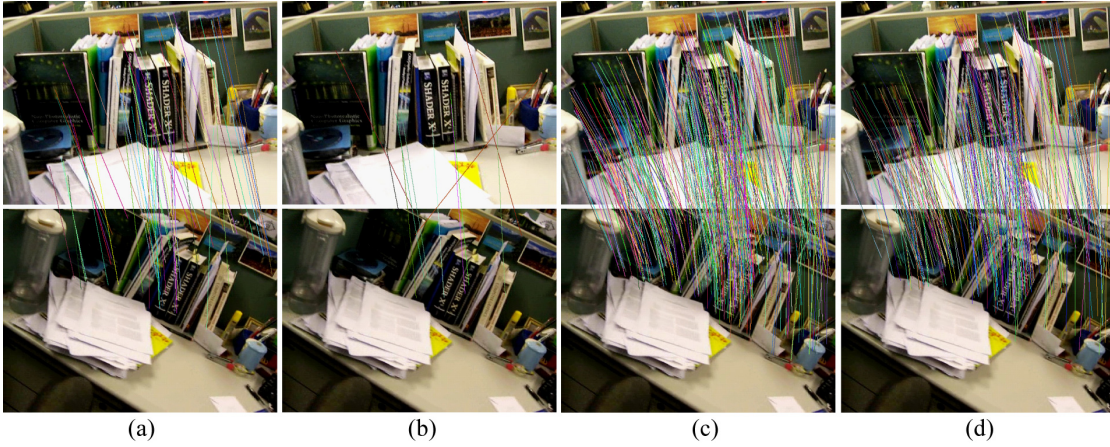


Fig. 2. Feature matching comparison. (a) First-pass matching by SIFT descriptor comparison. There are 958 features detected in the first image; but only 53 matches are found. (b) Additional matches by directly searching the correspondences along the epipolar lines by comparing the SIFT descriptors. Only 11 additional matches are found. (c) Second-pass matching by planar motion segmentation with outlier rejection. 399 (i.e. 53 + 346) matches are obtained. (d) Matching result of [51]. 314 matches are obtained.

Algorithm 2 Planar Motion Segmentation

1. Put all matches into a set Ω .
 2. For $k = 1, \dots, N_{\max}$, $\%N_{\max}$ is the maximum number of the homographies.
 - 2.1 Use RANSAC to estimate homography $H_{t,t+1}^k$ that has the maximum inliers.
 - 2.2 Remove the inliers from Ω . If the size of Ω is small enough, stop; otherwise, continue.
-

iteratively by taking the partial derivative w.r.t. \mathbf{x}_{t+1}^k and setting it to zero:

$$\frac{\partial S_{t,t+1}^k(\mathbf{x}_{t+1}^k)}{\partial \mathbf{x}_{t+1}^k} = 0. \quad (2)$$

$I_{t+1}(\mathbf{x} + \Delta\mathbf{x})$ is approximated by a Taylor expansion truncated up to its first order:

$$I_{t+1}(\mathbf{x} + \Delta\mathbf{x}) \approx I_{t+1}(\mathbf{x}) + g_{t+1}^\top(\mathbf{x}) \cdot \Delta\mathbf{x}, \quad (3)$$

where g_{t+1}^\top is the image gradient in the $(t+1)^{th}$ frame. With the computed gradients, we propose an iterative solver to optimize (1) by first initializing \mathbf{x}_{t+1} as the midpoint between $\hat{\mathbf{x}}_t^k$ and its projection to $l_{t,t+1}(\mathbf{x}_t)$, as shown in Fig. 3. Then we iteratively update \mathbf{x}_{t+1} by solving (2). In iteration $i+1$, \mathbf{x}_{t+1} is updated as

$$\mathbf{x}_{t+1}^{k,(i+1)} = \mathbf{x}_{t+1}^{k,(i)} + \Delta\mathbf{x},$$

where $\mathbf{x}_{t+1}^{k,(i)}$ denotes the value of \mathbf{x}_{t+1}^k in iteration i . This procedure continues until $\Delta\mathbf{x}$ is sufficiently small.

The found match is denoted as \mathbf{x}_{t+1}^k . With the set of homographies $\{H_{t,t+1}^k | k = 1, \dots, N\}$, we can find several matches $\{\mathbf{x}_{t+1}^k | k = 1, \dots, N\}$. Only the best one $j = \min_k \sum_{\mathbf{y} \in W} |\hat{I}_t^k(\hat{\mathbf{x}}_t^k + \mathbf{y}) - I_{t+1}(\mathbf{x}_{t+1}^k + \mathbf{y})|$ is kept.

In case the feature motion cannot be described by any homographies or feature correspondence is indeed missing, the found match is actually an outlier. We detect it with the

following conditions:

$$\begin{cases} \sum_{\mathbf{y} \in W} |\hat{I}_t^j(\hat{\mathbf{x}}_t^j + \mathbf{y}) - I_{t+1}(\mathbf{x}_{t+1}^j + \mathbf{y})| > \tau_c |W|; \\ d(\mathbf{x}_{t+1}^j, l_{t,t+1}(\mathbf{x}_t)) > \tau_e; \\ \|\hat{\mathbf{x}}_t^j - \mathbf{x}_{t+1}^j\| > \tau_h. \end{cases}$$

These conditions represent the constraints of color constancy, epipolar geometry and homography respectively. If any of them is satisfied, \mathbf{x}_{t+1}^j is treated as an outlier. τ_c is set to a small value (0.02 in our experiments) since the image is rectified. The remaining two parameters are $\tau_e = 2$ and $\tau_h = 10$. Considering points may not strictly undergo planar transformation, τ_h is set to a relatively large value.

Fig. 2(c) shows the result after the second-pass matching. Compared to our original scheme in [51] (Fig. 2(d)), our modified method does not need to perform additional KLT matching to refine the match position. It is thus more reliable and runs faster. The computation time is only 18ms with GPU acceleration. The number of credible matches also increases.

The two-pass matching can produce many long tracks. Each track has a group of descriptors. They are similar to each other in the same track due to the matching criteria. We compute the average of the descriptors over the track, and denote it as *track descriptor* $\mathbf{p}(\mathcal{X})$. It is used in the following non-consecutive track matching.

V. NON-CONSECUTIVE TRACK MATCHING

In this stage, we match features distributed in different subsequences, which is vital for drift-free SfM estimation. If we select all image pairs in a brute-force manner, the process can be intolerably costly for a long sequence. A better strategy is to estimate content similarity among different images first.

Our non-consecutive track matching (NCTM) method contains two steps. First, similarity of different images is coarsely estimated by constructing an $n \times n$ symmetric match matrix M , where n is the number of frames. $M(i, j)$ stores overlapping confidence between images I_i and I_j .

Fig. 4(a) shows the initially estimated match matrix for the “Desktop” sequence. Bright pixels are with high overlapping confidence where many common features should exist. In the second step, with this match matrix, we select the frame pairs with maximum overlapping confidence to perform feature matching, and update the match matrix iteratively. Matrix estimation and non-consecutive feature matching are benefitted from each other to dramatically simplify computation. Fig. 4(c) shows our finally estimated match matrix.

For speedup, we extract keyframes based on the result of consecutive feature tracking described in Section IV. Frame 1 is selected as the first keyframe. Then we select frame i as the second keyframe if it satisfies $N_1(1, i) \geq m_1$ and $N_1(1, i+1) < m_1$, where $N_1(i, j)$ denotes the number of common features between frames i and j . Other keyframes are selected as follows. For the two recent keyframes with indices i_1 and i_2 in the original sequence, we select frame j ($j > i_2$) as the new keyframe if it is the farthest one from i_2 that satisfies $\{N_1(i_1, j) \geq m_1, N_2(i_1, i_2, j) \geq m_2\}$, where $N_2(i_1, i_2, j)$ denotes the number of common points among the three frames (i_1, i_2, j) . This step is repeated until all frames are processed. In our experiments, $m_1 = 100 \sim 500$ and $m_2 = 50 \sim 300$. Without special notice, the following procedures are only performed on keyframes.

A. Fast Match Matrix Estimation

The first stage is to quickly produce a match matrix. We use a hierarchical K-means approach similar to [33] to cluster the track descriptors. The root contains all the descriptors. It is partitioned into b subgroups by the K-means method. Each sub-cluster consists of the descriptor vectors closest to the center. The same procedure is recursively applied to all subgroups and terminates when the variance of all descriptors in a final (leaf) cluster is less than a threshold. The leaf clusters provide a partition of all tracks. We measure the coarse overlapping confidence among non-consecutive frames by counting the number of corresponding tracks that are clustered into the same groups (depicted in Algorithm 3).

All elements in M are initialized to zeros. In each iteration of Algorithm 3, $M(i, j)$ is increased by 1 if two features respectively in frames i and j are in the same leaf node of the tree. With the objective of non-consecutive frame matching, we exclude the cases that two tracks in the same group span common frames (i.e., $f(\mathcal{X}_u) \cap f(\mathcal{X}_v) \neq \emptyset$) to exclude track self-matching. Hence, the diagonal band of our matrix has no value. For further acceleration, we generally only select long tracks that span 5 or more keyframes to estimate overlapping confidence. In our experiments, for the “Desktop” sequence, the match matrix estimation only takes 1.08 second, with a total of 5,935 selected feature tracks.

Our method can robustly handle dense image sequences, unlike FABMAP [9] that assumes sparsely sampled ones. When applying FABMAP to “Desktop” sequence, it fails to detect any loops. Differently, we manually sample the

Algorithm 3 Match Matrix Estimation

1. Initialize M as a zero matrix.
 2. For each track cluster G_k ($k = 1, \dots, K$), % K is the number of the final clusters.
 - For any $(\mathcal{X}_u, \mathcal{X}_v)$ in G_k , if $f(\mathcal{X}_u) \cap f(\mathcal{X}_v) = \emptyset$,
 for any $i \in f(\mathcal{X}_u)$ and $j \in f(\mathcal{X}_v)$,
 $M(i, j) += 1$,
 $M(j, i) += 1$.
-

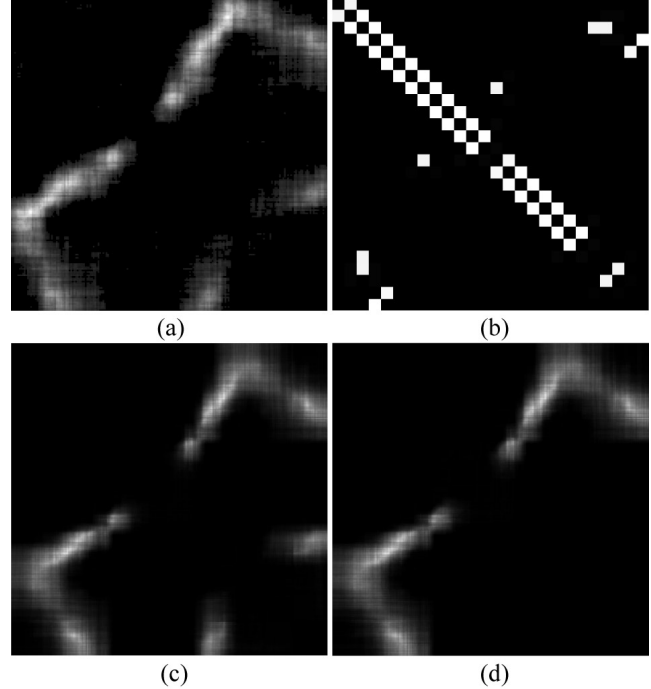


Fig. 4. Match matrix estimation for the “Desktop” sequence containing 941 frames. (a) Our initially estimated match matrix based on the keyframes. The matrix size is scaled for better visualization. (b) The estimated match matrix by FABMAP [9] on the re-sampled sequence that contains 26 frames. The matrix size is scaled for better visualization. (c) The final estimated match matrix (for all frames) after our non-consecutive matching based on (a). (d) The final estimated match matrix (for all frames) after our non-consecutive matching based on (b).

original sequence until the common points between adjacent sampled frames are no more than 100 (i.e., 26 sampled frames). As shown in Fig. 4(b), a few overlapping image pairs are identified; but they are still insufficient to match many common features.

B. Non-Consecutive Track Matching

We propose a new matching strategy to match the common tracks distributed in different subsequences. The number of common features between two frames can be coarsely reflected by the estimated match matrix M described in Section V-A, which, however, is not enough for precisely finding the matching images. Obtaining accurate common features number requires exhaustively matching each frame pairs. Fortunately, by utilizing the video continuity, an approximate estimate of common features number in a frame pair can be quickly achieved according to the

matched tracks in their neighboring frames. After matching a frame pair (t_1^0, t_2^0) , the set of matched track pairs $C_{\mathcal{X}} = \{(\mathcal{X}_1, \mathcal{X}_2)\}$ can approximately represent the number of common tracks for the neighboring frame pairs. The matched track pairs in frame pair (t_1, t_2) can be expressed as

$$C_{\mathcal{X}}(t_1, t_2) = \{(\mathcal{X}_1, \mathcal{X}_2) | t_1 \in f(\mathcal{X}_1), t_2 \in f(\mathcal{X}_2), (\mathcal{X}_1, \mathcal{X}_2) \in C_{\mathcal{X}}\}. \quad (4)$$

The number of common features in (t_1, t_2) can be approximated by $|C_{\mathcal{X}}(t_1, t_2)|$ as long as (t_1, t_2) shares sufficient common tracks with (t_1^0, t_2^0) . Therefore, we maintain an updating match matrix M^* , computed as

$$M^*(t_1, t_2) = |C_{\mathcal{X}}(t_1, t_2)|, \quad (5)$$

to propagate the overlapping confidence from (t_1^0, t_2^0) toward the neighboring frame pairs, and determine where the next matching should be performed. Details are given below.

Main Procedure: As illustrated in our supplementary video, we first detect the brightest point (t_1^0, t_2^0) in the estimated match matrix M . The value of $M(t_1^0, t_2^0)$ is also denoted as M_{\max} . If $M(t_1^0, t_2^0)$ is larger than a threshold, several common features may exist. After matching (t_1^0, t_2^0) , we collect and put the matched track pairs into $C_{\mathcal{X}}$ and update M^* according to Eq. (5). In particular, we set $M^*(t_1^0, t_2^0) = 0$, indicating (t_1^0, t_2^0) is matched. Next, we repeatedly select the brightest point (t_1^k, t_2^k) in the updating matrix M^* , match (t_1^k, t_2^k) , and update $C_{\mathcal{X}}$ and M^* accordingly. This procedure continues until $M^*(t_1^k, t_2^k) < 50$. Then we go to another region by re-detecting the brightest point in M that has not been processed. The step ends if the brightest value is smaller than $0.1M_{\max}$.

Frame pair matching: When entering a new bright region, we perform the classical 2NN matching for (t_1^0, t_2^0) . Then each matching pair (t_1^k, t_2^k) is detected from the updating matrix M^* . Thus there are $M^*(t_1^k, t_2^k)$ common features found previously. We use these matches to estimate the fundamental matrix $F_{t_1^k, t_2^k}$ of frame pair (t_1^k, t_2^k) , and re-match those outlying features along the epipolar lines. We further search the correspondences for other unmatched features along epipolar lines. This strategy not only speeds up the system but also obtains reliable feature matches.

Outlier rejection: Along with the fundamental matrix estimation between t_1^k and t_2^k , these $M^*(t_1^k, t_2^k)$ matches are classified into inliers and outliers. Since only part of matches are used to estimate $F_{t_1^k, t_2^k}$, the estimated $F_{t_1^k, t_2^k}$ could be biased. So we do not reject outliers immediately. Fortunately, each matched track pair $(\mathcal{X}_1, \mathcal{X}_2)$ undergoes multi-pass epipolar verification during processing the whole bright region.

We record all the verification results for each $(\mathcal{X}_1, \mathcal{X}_2)$, and determine inliers/outliers after all bright regions are processed. Suppose $(\mathcal{X}_1, \mathcal{X}_2)$ is classified as an inlier match N_I times and as an outlier match N_O times. We reject $(\mathcal{X}_1, \mathcal{X}_2)$ if $N_I < s \cdot N_O$ ($s = 1 \sim 4$ in our experiments).

In addition, we use the following strategy to remove the potential matching ambiguity. For example, a track \mathcal{X}_1 may find two corresponding tracks \mathcal{X}_2 and \mathcal{X}_2' , where \mathcal{X}_2 and \mathcal{X}_2' have overlapping frames. So the track matches $(\mathcal{X}_1, \mathcal{X}_2)$ and $(\mathcal{X}_1, \mathcal{X}_2')$ conflict with each other. In this case, we simply select the best match with the largest N_I , and regard the other as an outlier.

Benefits: The proposed matching method outperforms previous ones in the following aspects. In our original method proposed in [51], a rectangular region in the roughly estimated match matrix M is sought each time and local exhaustive track matching is performed for all frame pairs in it. It could involve a lot of unnecessary matching for non-overlapping frames and repeated feature comparison. Our current scheme only selects the frame pairs with sufficient overlapping, and matches each pair of frames and most tracks at most once.

Standard image matching is to find a set of most similar images given the query one. This scheme has been extensively used in large-scale SfM [2], [13] and realtime SLAM systems for loop closure detection [4], [6], [24]. It, however, also may involve unnecessary matching for unrelated frame pairs and miss those with considerable common features. It is because image similarity based on appearance may not be sufficiently reliable. In contrast, we progressively expand frames with track matching. The expansion is not fully related to the match matrix. Therefore a very rough matrix is enough to give a good starting point. Practically, as long as there is one good position, our system can extend it to the whole overlapping region accurately. To verify this, we provided two refined match matrices based on two different rough match matrices, as shown in Figs. 4(a) and (b). Although the two initially estimated match matrices are different and only based on keyframes, the finally estimated match matrices after our non-consecutive track matching are quite similar (except the bottom right area, where the initial match matrix by FABMAP does not provide any high confidence elements), which demonstrates the effectiveness of our method.

C. SfM for Multiple Sequences

In a large-scale scene, we generally take a very long sequence or multiple sequences from geographically different regions. For a very long sequence, the accumulated reconstruction error might be too large to be eliminated by a traditional bundle adjustment algorithm. It is natural to divide the long sequence into multiple short parts to alleviate error accumulation. For multiple sequences, how to efficiently match and register them in a unified 3D system was not extensively discussed in previous work. In our feature tracking system, this goal can be naturally accomplished.

Given one or multiple sequences, we first split long ones, making each new sequence generally contains only 1000 \sim 3001 frames. The split neighboring sequences have overlapping frames for reliable matching. The sequence set is denoted as $\{V_i | i = 1, \dots, n\}$. Then we apply feature

tracking to each V_i , and estimate its 3D structure and camera motion using a SfM technique similar to that of [52]. The major modification is that we use known intrinsic camera parameters, and simply select an initial frame pair that has sufficient matches and a large baseline to start SfM. For each sequence pair, we use the algorithm described in Section V-A to rapidly estimate the match matrix such that related frames in any two different sequences can be found and common features can be matched by the algorithm introduced in Section V-B. This method quickly yields a matching graph, whose nodes represent different sequences, and each edge denotes a set of common feature tracks between two sequences. We finally use the segment-based SfM method described in Section VI to efficiently recover and globally register 3D structures and camera trajectories, as shown in Fig. 1(b).

VI. SEGMENT-BASED COARSE-TO-FINE SFM

With the independently reconstructed sequences and matched common tracks, we align them in a unified 3D coordinate system. For a long loopback sequence, error accumulation could be serious, making traditional bundle adjustment easily stuck in local optima. It is because the first a few iterations of bundle adjustment aggregate accumulation errors at the joint loop points, which are hard to be propagated to the whole sequence.

We propose splitting each sequence into multiple segments – each is with a similarity transformation. Only these transformations and overlapping points across different segments are optimized. We name it segment-based refinement and illustrate it in Fig. 6. Although the method of [24], which performs global adjustment by clustering frames into multiple disjoint sets, is conceptually similar to ours, it cannot be applied to our datasets. For example, when the camera moves in a large-scale scene, the recovered camera trajectory may not be perfectly closed (see our supplementary video). The geodesic-distance-based segmentation to cluster frames could make inconsistent structure be put into a single body, complicating alignment-error reduction. Local optimization within each body cannot address this issue, as errors are caused by global misalignment. Our new split scheme is as follows.

In the beginning, we order all sequences and define the one that contains the maximum number of tracks merged with others as the reference. Without losing generality, we define it as sequence #1, denoted as V_1 . Its local 3D coordinate system is also set as the reference. Then with the common tracks among different sequences, we can estimate the coordinate transformation for each sequence j (i.e., V_j), denoted as $T_j = (s_j, R_j, t_j)$, where s_j is the scale factor, R_j is the rotational matrix, and t_j is the translation vector. For the reference sequence, s_1 have value 1, R_1 is an identity 3×3 matrix, and $t_1 = (0, 0, 0)^T$.

Instead of directly optimizing camera motion for frames, we optimize the transformation among frame segments. Each segment is assigned with a similarity transformation, and the relative camera motion between frames in each

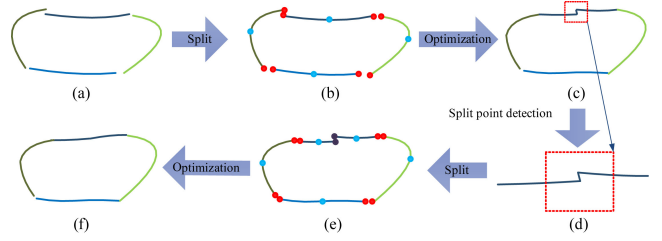


Fig. 6. Segment-based coarse-to-fine refinement. (a) Recovered camera trajectories marked with different colors. (b) Each sequence is split into 2 segments where endpoints and midpoints are highlighted. (c) Refined camera pose after the first iteration, where errors are redistributed. (d) “Split points”, which are joints of largely inconsistent camera motion for neighboring frames. (e) Sequence separation by split points. (f) Refined camera trajectories after 2 iterations.

segment is fixed, so that the number of variables is small enough for efficient optimization. Different from [24], which clusters frames using geodesic distances, we propose clustering neighboring and geometrically consistent frames into segments. The position at which two consecutive frames are inconsistent is defined as a “split point”.

In order to detect them, the most straightforward way is to project the common points between each consecutive frame pair into the two images and check the re-projection errors. However, in the regions where different sequences overlap, the re-projection error is always larger than other regions. As a result, excessive split points would be detected in the overlapping regions. Splitting most of such frame pairs does not help reduce the overall re-projection error because the relative motion here may not be worse than other regions. We explicitly find split points such that the re-projection error is most likely to be reduced. Assume each frame k is associated with a small similarity transformation T_k , which is parameterized as a 7-vector a_k (three Rodrigues components for rotation, 3D translation and a scaling variable). If we minimize the re-projection error w.r.t. a_k , the steepest descent direction is

$$g_k = \sum_{i=1 \dots N_k} A_i^T e_i, \quad (6)$$

where N_k is the number of points visible in frame k , and A_i is the Jacobian matrix $A_i = \partial \pi(P_k X_i) / \partial a_k$. π is the projection function. e_i is the re-projection error $e_i = \mathbf{x}_i - \pi(P_k X_i)$, which is reduced along the direction of g_k . For two consecutive frames, if e_i can be reduced along the same direction, they are associated with the same similarity transformation. Otherwise, a split point is placed there. The inconsistency between two consecutive frames is defined as the angle between the two steepest descent directions

$$C(k, k+1) = \arccos \frac{g_k^T \cdot g_{k+1}}{\|g_k\| \cdot \|g_{k+1}\|}. \quad (7)$$

For verification, we group every 100 consecutive frames into one segment for the “Desktop” (Fig. 5(a)), and apply a certain transformation to each segment (Fig. 5(b)). As expected, the re-projection errors distribute in the whole overlapping regions. In contrast, the angle between the

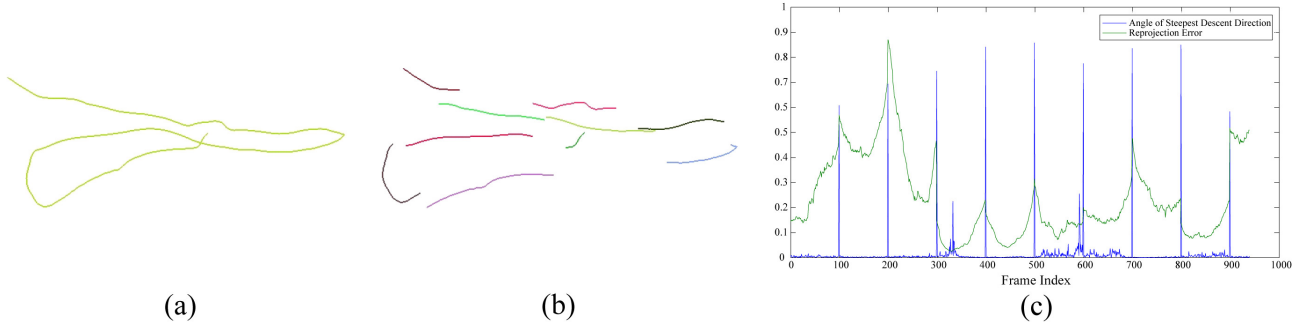


Fig. 5. Split point detection. (a) Original camera trajectory of the “Desktop” sequence. (b) Splitted camera trajectories. Each segment contains 100 frames. (c) Re-projection errors (green curve) and angles of steepest descending direction (blue curve). Values are all normalized to [0,1]. The angle more accurately reflects the split result quality compared to the re-projection error.

steepest descent directions reliably reflects the splitting result.

We progressively segment the sequences. At the t^th iteration, each sequence is divided into 2^t segments. We compute $C(k, k+1)$ for all k and detect the $2^t - 1$ split points with largest $C(k, k+1)$. We put the consecutive frames in between two adjacent split points into a segment, and use the method described as follows to estimate the similarity transformations and submaps jointly for all segments. When the optimization is done, we detect split points for each sequence again, and re-separate the sequence into multiple segments. We repeat this process until the average reprojection error is below 1 pixel or each segment contains only one frame. Errors are progressively propagated and reduced. The procedure of our segment-based coarse-to-fine refinement scheme is illustrated in Fig. 6.

Algorithm Details: Suppose the number of detected split points among all n sequences is m . We break the sequences into a total of $n' = 2(n + m)$ segments. Each of them is with a similarity transformation $T_j^w = (s_j^w, R_j^w, t_j^w)$, where $j = 1, \dots, n'$, w.r.t. the world coordinate. We use the bundle adjustment (BA) to refine the reconstructed 3D feature points (i.e., submaps) with these similarity transformations. Different from traditional BA, the camera parameters inside each segment are fixed, we thus only update the similarity transformation. The procedure is to first transform one 3D point in the world coordinate to a local one with parameters T^w . Then traditional perspective camera projection is employed to compute the re-projection error. Our BA function is written as

$$\min \sum_{i=1}^{N'} \sum_{j=1}^{n'} \sum_{k=1}^{n_j} w_{i,j,k} \|\pi(K_{j,k}(R_{j,k}(s_j^w R_j^w X_i + t_j^w) + t_{j,k})) - \mathbf{x}_{i,j,k}\|^2, \quad (8)$$

where n_j is the number of frames in the j -th subsequence, N' is the number of the 3D feature points, and n' is the number of the segments. π is the projection function. $\mathbf{x}_{i,j,k}$ is the image location of X_i in the k -th frame of the j -th subsequence. $K_{j,k}$, $R_{j,k}$, and $t_{j,k}$ are the intrinsic matrix, rotational matrix, and translation vector, respectively. $w_{i,j,k}$ is defined as

$$w_{i,j,k} = \begin{cases} 1, & \text{If point } i \text{ is visible in frame } k \text{ in sequence } j \\ 0, & \text{Otherwise} \end{cases}$$

We use Schur complement and forward substitution [44] to solve the normal equation, which separates the updating of rigid transformation and of 3D points in each iteration. It reduces the large linear system to a linear symmetric one with scale $6n' \times 6n'$ for updating transformation. It makes 3D point estimation much cheaper because each point can be updated independently by solving a 3×3 linear symmetric system. Moreover, since only a few segment pairs share common points, the Schur complement is rather sparse.

It is notable that in [25], the system of Schur complement was explicitly constructed and solved by Cholesky decomposition. Wu et al. [48] implicitly built the Schur complement for parallel computing. They did not take full advantage of the sparsity property. In contrast, we solve the problem by efficient preconditioned conjugate gradient similar to that of [48], utilizing sparsity of the matrices. This strategy significantly reduces computation (about 3 times faster in our experiments).

Because the size of the linear system is actually determined by n' . We estimate n' based on the available memory. Once the size n' linear system is reached, SfM refinement is performed in the following two steps. In the first one, we only select the $m = n'/2 - n$ split points with maximal $C(k, k+1)$ to split the sequences, and solve Eq. (8) to refine the result. In the second step, we perform a local bundle adjustment for each sequence j iteratively by re-splitting sequence j to multiple segments with detected split points and refining them by solving Eq. (8) while fixing cameras and 3D points in other sequences. This process stops when all sequences are processed. This strategy makes it possible to efficiently and robustly handle large data with limited memory consumption.

Finally, we fix the 3D points and estimate the camera motion respectively for all frames. During the course of iteration, errors are quickly reduced. Please refer to our supplementary video² to understand how our method works.

²<http://www.cad.zju.edu.cn/home/gfzhang/projects/tracking/featuretracking/featuretracking.wmv>

TABLE II
RUNNING TIME.

Datasets	Frames	Step	Resolution	Feature Tracking		SfM Estimation			Propagation	
				Cons.	Non-Cons.	Submap Est.	Align	Refine	Feature Propagation	Camera Est.
Desktop	941	1	640×480	46.5s	5.8s	14.1s	-	-	-	-
Circle	2,129	3	960×540	63.1s	40.9s	13.8s	-	-	13.9s	4.2s
Street	22,799	5	960×540	7.4 min.	7.5 min.	176.0s	3.6s	32.0s	3.4 min.	65.1s
Garden	95,476	3 or 5	960×540	27.4 min.	31.2 min	588.1s	2.5s	130.4s	15.6 min.	3.7 min.

VII. EXPERIMENTAL RESULTS

We evaluate our method on several challenging sequences. Running time is listed in Table II excluding I/O, which is obtained on a desktop PC with an Intel i7-4770K CPU, 8GB memory, and a NVIDIA GTX780 graphics card. The operating system is 64-bit Windows 7. Only the feature tracking component is accelerated by GPU³. For SfM estimation, we optimize the code by applying SSE instructions, but only use a **single thread** without parallel computing. For the sequences captured by us, since the intrinsic matrix is known, we optimize the SfM code by incorporating this prior to improve the robustness and efficiency.

As our consecutive point tracking can handle wide-baseline images, frame-by-frame tracking is generally unnecessary. In our experiments, the system generally extracts one frame for every 3 ~ 5 frames to apply feature tracking. We quickly propagate the feature points to other frames by KLT with GPU acceleration. This trick further saves computation. In addition, in order to reduce image noise and blur, for each input frame I_t , we perform matching with two past frames. One is the last frame I_{t-1} , and the other (denoted as $I_{t'}$) is the farthest frame that shares over 300 common features with I_{t-1} . Note that only a small number of features in $I_{t'}$ need to be matched with I_t , which does not increase computation much.

A. Quantitative Evaluation

We compare the feature tracking methods of consecutive SIFT matching (C-SIFT), our consecutive point tracking (CPT), brute-force SIFT matching (BF-SIFT), our consecutive point tracking with non-consecutive track matching (CPT+NCTM), our consecutive point tracking with keyframe-based non-consecutive track matching (CPT+KNCTM).

C-SIFT extracts and matches SIFT features only in consecutive frames. It is a common strategy for feature tracking. The advantage is that the complexity is linear to the number of frames. However, feature dropout could occur due to global indistinctiveness or image noise, which causes producing many short tracks. Examples are included in our supplementary material. The brute-force SIFT matching exhaustively compares extracted SIFT features, whose complexity is quadratic to the number of processed frames. In

TABLE III
PERFORMANCE OF DIFFERENT ALGORITHMS.

Algorithms	Running Time	Average Track Length
C-SIFT	38.4s	1.73
CPT	63.1s	2.28
CPT+NCTM	150.7s	3.10
CPT+KNCTM	104.0s	2.68
BF-SIFT	1086.4s	2.71

comparison, the complexity of our method (CPT+NCTM) is almost linear to the frame number while high quality results are guaranteed.

The “Circle” sequence contains 2,129 frames. To make computation feasible for a few prior methods, we select one frame for every 3 consecutive ones, which forms a new sequence containing 710 frames in total. Table III lists the running time with GPU acceleration. Our consecutive point tracking (CPT) needs a bit more time than C-SIFT. But it significantly extends the lifetime of most tracks, as shown in the track length histogram contained in our supplementary material. With our non-consecutive track matching, common feature tracks scattered over disjoint subsequences are connected, further expanding track lifetime. Compared with the computationally most expensive BF-SIFT, our result (CPT+NCTM) obtains more long feature tracks and the method is about 7 times faster. With keyframe-based acceleration, our non-consecutive track matching time becomes much faster (from 87.6s to 40.9s), without influencing much matching result. Table III lists the average length of tracks (track length ≥ 1)⁴. The quality of SfM computed by BF-SIFT, CPT+NCTM and CPT+KNCTM are quite comparable, as shown in our supplementary material.

B. Comparison with VisualSFM and PTAMM

We compare our ENFT-SFM system with the recent state-of-the-art work VisualSFM [47], using the “Street” data. This data set contains 2 long sequences with 22,799 frames in total. One camera moves along a street in two rounds, and another camera moves along the street on the other side. We split them to 18 shorter sequences, and extract one frame for every 5 frames to constitute new keyframe sequences for feature tracking and camera motion estimation. For fair comparison, the input feature matching

³We use 64D descriptors for SIFT features. Our SIFT GPU implementation is inspired by [45] but runs faster.

⁴The computed average track length is short because we also take into account of the unmatched features (i.e. track length = 1).

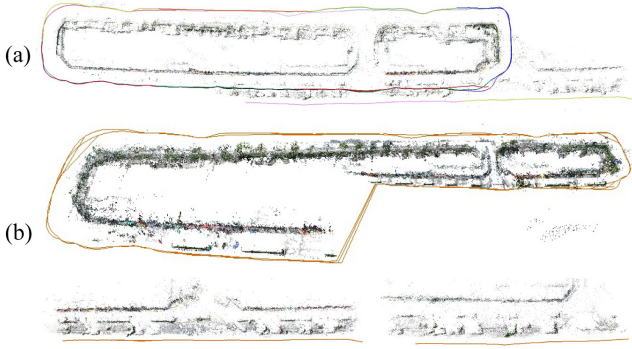


Fig. 7. SfM comparison on the “Street” example. (a) SfM result of ENFT-SfM. (b) SfM result of VisualSfM, which is separated to 3 individual models.

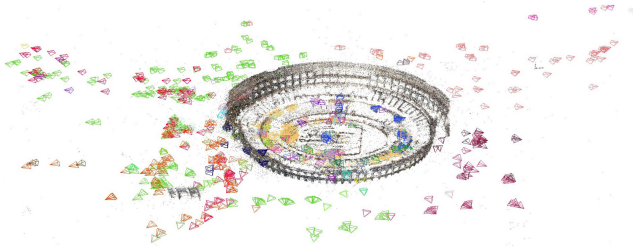


Fig. 8. “Colosseum” example. Cameras in the same sequence are with the same color.

results are all obtained by the proposed feature tracking algorithm.

We found VisualSfM does not work well on these long sequences – the produced SfM result has the drift problem and the whole scene is separated into 3 individual models, as shown in Fig. 7(b). One possible reason is that VisualSfM employs incremental SfM, which may not effectively eliminate accumulated errors as our segment-based SfM. Another possible reason is that VisualSfM does not sufficiently exploit sequence continuity/ordering, and heavily relies on good feature matches⁵. Our SfM estimation without parallel computing only takes 212 seconds for this example, which is even faster than VisualSfM enabling GPU (1,089 seconds).

We also compare our system with the well-known SLAM system PTAMM [20], [21], [4], which is designed for working in a relatively small space. Even for the smallest “Desktop” sequence, our system still outperforms PTAMM. Please refer to our supplementary video for the comparison.

C. Results on General Image Collections

Although our system is originally designed for handling sequences, it can be naturally extended to work with general image collections. The basic idea is to separate the unordered image data to a set of sequences according to their common matches.

We first select two images with the maximum number of common features to constitute an initial sequence. Then we



Fig. 9. Large-baseline matching may not be handled by our method. The first-pass matching on an image pair only obtains 9 matches. The third image is the incorrect rectification result.

select another image, which has the most common features with the head or tail frame, and add it into the sequence as the new head or tail. This process repeats until no image can be added. Then we begin to build another sequence based on remaining images. For some 3D points that have only one or two corresponding features in one sequence, we additionally select related images from other sequences to help estimate the 3D positions.

Figure 8 shows our SfM result on Colosseum data [46], which contains 1,164 images. We use the same feature matching method [45] of VisualSfM to match the images. Because our current SfM implementation requires that the intrinsic camera parameters and radial distortion are known for each image, we calibrate the matched feature positions according to the calibrated parameters by VisualSfM. Then we use our extended segment-based SfM method to estimate camera poses and 3D points. The processing time of our SfM estimation in a single thread is 125 seconds, which is shorter than that of VisualSfM enabling GPU (269 seconds), and much shorter than that of VisualSfM disabling GPU (about 2,864 minutes).

D. Parameter Configuration and Limitation

The parameters can be easily set in our system because most of them are not sensitive and use default values. The most important parameter is τ_c , which controls the strength to mark outliers. A large τ_c could result in many matches, and also introduces outliers. In our experiments, we conservatively set τ_c to a small value 0.02. By removing a small set of matches, the system becomes reliable for high-quality SfM. Please refer to our supplementary material⁶ for result comparison using different τ_c .

The proposed two-pass matching works best if the scene can be represented by many planes. For a video sequence with dense frames, this condition can be generally achieved because image transformation between two consecutive frames is small for viable approximation by one or multiple homographies. We note even if the scene deviates from piecewise planarity, our second-pass matching still works as rectified images are close to the target ones. Our method may be not suitable for wide-baseline sparse images where

⁵The used VisualSfM has been modified by the author to make special optimization for our data to alleviate the drift problem.

⁶<http://www.cad.zju.edu.cn/home/gfzhang/projects/tracking/featuretracking/supplement.pdf>

the number of matches by first-pass matching is too small. Fig. 9 shows a failure example. The first pass process only obtains 9 matches. So no appropriate homography can be estimated.

VIII. CONCLUSION AND DISCUSSION

We have presented a robust and efficient non-consecutive feature tracking (ENFT) method for robust SfM, which consists of two main steps, i.e., consecutive point tracking and non-consecutive track matching. Different from typical sequential matchers, e.g., KLT, we use invariant features and propose a two-pass matching strategy to significantly extend the track lifetime and reduce the feature sensitivity to noise and image distortion. The obtained tracks avail estimating a match matrix to detect disjointed subsequences with overlapping views. A new segment-based coarse-to-fine SfM estimation scheme is also introduced to effectively reduce accumulation error for long sequences. The presented ENFT-SfM system can handle tracking and registering large video datasets with small memory consumption.

Our ENFT method greatly helps SfM, and considers feature tracking on non-deforming objects by tradition. Part of our future work is to handle dynamic objects. Also, although the proposed method is based on SIFT features, there is no limitation to use other representations, e.g., SURF or other cheaper features, for further acceleration.

ACKNOWLEDGEMENTS

We would like to thank Changchang Wu for his kind help in running VisualSfM with our datasets. This work is partially supported by National Science and Technology Support Plan Project (No. 2012BAH35B02), NSF of China (Nos. 61232011 and 61272048), the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20110101130011), the Fundamental Research Funds for the Central Universities (2015XZZX005-05), and a Foundation for the Author of National Excellent Doctoral Dissertation of PR China (No. 201245).

REFERENCES

- [1] Sameer Agarwal, Noah Snavely, Steven M. Seitz, and Richard Szeliski. Bundle adjustment in the large. In *ECCV (2)*, pages 29–42, 2010.
- [2] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. Building rome in a day. In *ICCV*, pages 72–79, 2009.
- [3] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.
- [4] Robert O. Castle, Georg Klein, and David W. Murray. Video-rate localization in multiple maps for wearable augmented reality. In *ISWC*, pages 15–22, 2008.
- [5] Laura A. Clemente, Andrew J. Davison, Ian D. Reid, José Neira, and Juan D. Tardós. Mapping large loops with a single hand-held camera. In *Robotics: Science and Systems*, 2007.
- [6] Brian Clipp, Jongwoo Lim, Jan-Michael Frahm, and Marc Pollefeys. Parallel, real-time visual slam. In *IROS*, pages 3961–3968, 2010.
- [7] Kai Cordes, Oliver Muller, Bodo Rosenhahn, and Jorn Ostermann. Feature trajectory retrieval with application to accurate structure and motion recovery. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Song Wang, Kim Kyungnam, Bedrich Benes, Kenneth Moreland, Christoph Borst, Stephen DiVerdi, Chiang Yi-Jen, and Jiang Ming, editors, *Advances in Visual Computing*, volume 6938 of *Lecture Notes in Computer Science*, pages 156–167. Springer Berlin / Heidelberg, 2011.
- [8] David J. Crandall, Andrew Owens, Noah Snavely, and Dan Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *CVPR*, pages 3001–3008, 2011.
- [9] Mark Cummins and Paul M. Newman. Appearance-only slam at large scale with fab-map 2.0. *I. J. Robotic Res.*, 30(9):1100–1123, 2011.
- [10] Ethan Eade and Tom Drummond. Unified loop closing and recovery for real time monocular slam. In *BMVC*, 2008.
- [11] Chris Engels, Friedrich Fraundorfer, and David Nistér. Integration of tracked and recognized features for locally and globally robust structure from motion. In *VISAPP (Workshop on Robot Perception)*, pages 13–22, 2008.
- [12] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.
- [13] Jan-Michael Frahm, Pierre Fite Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, and Svetlana Lazebnik. Building rome on a cloudless day. In *ECCV (4)*, pages 368–381, 2010.
- [14] Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. Towards internet-scale multi-view stereo. In *CVPR*, pages 1434–1441, 2010.
- [15] Michael Grabner and Horst Bischof. Extracting object representations from local feature trajectories. In *Joint Hungarian-Austrian Conference on Image Processing and Pattern Recognition*, 2005.
- [16] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.
- [17] Michal Havlena, Akihiko Torii, and Tomás Pajdla. Efficient structure from motion by graph optimization. In *ECCV (2)*, pages 100–113, 2010.
- [18] Kin Leong Ho and Paul M. Newman. Detecting loop closure with scene sequences. *International Journal of Computer Vision*, 74(3):261–286, 2007.
- [19] Arnold Irschara, Christopher Zach, Jan-Michael Frahm, and Horst Bischof. From structure-from-motion point clouds to fast location recognition. In *CVPR*, 2009.
- [20] Georg Klein and David W. Murray. Parallel tracking and mapping for small ar workspaces. In *ISMAR*, pages 225–234, 2007.
- [21] Georg Klein and David W. Murray. Improving the agility of keyframe-based slam. In *ECCV (2)*, pages 802–815, 2008.
- [22] Kurt Konolige and Motilal Agrawal. Frameslam: From bundle adjustment to real-time visual mapping. *IEEE Transactions on Robotics*, 24(5):1066–1077, 2008.
- [23] Xiaowei Li, Changchang Wu, Christopher Zach, Svetlana Lazebnik, and Jan-Michael Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV (1)*, pages 427–440, 2008.
- [24] Jongwoo Lim, Jan-Michael Frahm, and Marc Pollefeys. Online environment mapping. In *CVPR*, pages 3489–3496, 2011.
- [25] M.I. A. Lourakis and A.A. Argyros. SBA: A Software Package for Generic Sparse Bundle Adjustment. *ACM Trans. Math. Software*, 36(1):1–30, 2009.
- [26] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [27] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981.
- [28] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vision Comput.*, 22(10):761–767, 2004.
- [29] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005.
- [30] Jean-Michel Morel and Guoshen Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM J. Img. Sci.*, 2(2):438–469, 2009.

- [31] Kai Ni, Drew Steedly, and Frank Dellaert. Out-of-core bundle adjustment for large-scale 3d reconstruction. In *ICCV*, pages 1–8, 2007.
- [32] David Nistér, Oleg Naroditsky, and James R. Bergen. Visual odometry. In *CVPR (1)*, pages 652–659, 2004.
- [33] David Nistér and Henrik Stewénus. Scalable recognition with a vocabulary tree. In *CVPR (2)*, pages 2161–2168, 2006.
- [34] Marc Pollefeys, David Nistér, Jan-Michael Frahm, Amir Akbarzadeh, Philippos Mordohai, Brian Clipp, Christoph Engels, David Gallup, Seon Joo Kim, Paul Merrell, C. Salmi, Sudipta N. Sinha, B. Talton, Liang Wang, Qingxiong Yang, Henrik Stewénus, Ruigang Yang, Greg Welch, and Herman Towles. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision*, 78(2-3):143–167, 2008.
- [35] Eric Royer, Maxime Lhuillier, Michel Dhome, and Jean-Marc Lavest. Monocular vision for mobile robot localization and autonomous navigation. *International Journal of Computer Vision*, 74(3):237–260, 2007.
- [36] Frederik Schaffalitzky and Andrew Zisserman. Automated location matching in movies. *Computer Vision and Image Understanding*, 92(2-3):236–264, 2003.
- [37] Grant Schindler, Matthew Brown, and Richard Szeliski. City-scale location recognition. In *CVPR*, 2007.
- [38] Jianbo Shi and Carlo Tomasi. Good features to track. In *CVPR*, pages 593–600, 1994.
- [39] Sudipta N. Sinha, Drew Steedly, and Richard Szeliski. Piecewise planar stereo for image-based rendering. In *ICCV*, pages 1881–1888, 2009.
- [40] Sudipta N. Sinha, Drew Steedly, and Richard Szeliski. A multi-stage linear approach to structure from motion. 2010.
- [41] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.
- [42] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Skeletal sets for efficient structure from motion. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [43] Drew Steedly, Irfan A. Essa, and Frank Dellaert. Spectral partitioning for structure from motion. In *ICCV*, pages 996–1003, 2003.
- [44] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Workshop on Vision Algorithms*, pages 298–372, 1999.
- [45] Changchang Wu. SiftGPU: A GPU implementation of scale invariant feature transform (SIFT). <http://cs.unc.edu/ccwu/siftgpu>, 2007.
- [46] Changchang Wu. Towards linear-time incremental structure from motion. In *3DV*, 2013.
- [47] Changchang Wu. Visualsfm: A visual structure from motion system. <http://homes.cs.washington.edu/ccwu/vsfm/>, 2013.
- [48] Changchang Wu, Sameer Agarwal, Brian Curless, and Steven M. Seitz. Multicore bundle adjustment. In *CVPR*, pages 3057–3064, 2011.
- [49] Changchang Wu, Brian Clipp, Xiaowei Li, Jan-Michael Frahm, and Marc Pollefeys. 3d model matching with viewpoint-invariant patches (vip). In *CVPR*. IEEE Computer Society, 2008.
- [50] Christopher Zach, David Gallup, and Jan michael Frahm. Fast gain-adaptive klt tracking on the gpu. In *CVPR Workshop on Visual Computer Vision on GPU's (CVGPU)*, 2008.
- [51] Guofeng Zhang, Zilong Dong, Jiaya Jia, Tien-Tsin Wong, and Hujun Bao. Efficient non-consecutive feature tracking for structure-from-motion. In *ECCV (5)*, pages 422–435, 2010.
- [52] Guofeng Zhang, Xueying Qin, Wei Hua, Tien-Tsin Wong, Pheng-Ann Heng, and Hujun Bao. Robust metric reconstruction from challenging video sequences. In *CVPR*, 2007.